A Self-Tuned Thermal Compensation System for Reducing Process Variation Influence in Side-Channel Attack Resistant Dual-Rail Logic

Wei He, Marc Stottinger

Eduardo de la Torre, Veronica Diaz

I. Introduction

In modern cryptography, data is protected by utilizing strong cipher algorithms, which inevitably draws into numerous vulnerabilities from the implementations. Stunted by complex cipher systems, pure mathematic cryptanalysis became far to be viable when attacking modern crypto algorithms. In contrast to conventional cryptanalyses, side-channel attacks get around of the unaffordable mathematic computation, and specially excavate the decipherable physical phenomenons for retrieving the secrets [1]. The importance of side-channel attack resides on the fact that any logic elements in a complex system have unique physical features, like the toggling format of a transistor that is relying on the processed data (*i.e.*, data-dependent). By making the dependence computation between the predicted leakage for all possible values and the real-measured leakages, the right values can be revealed.

One typical defending strategy against side-channel analysis is to smoothen the data-dependent variations. Following this principle, a dual-rail architecture is utilized where the single rail in a normal circuit is replaced by two parallel rails. The two rails wok in the complementary fashion to balance the power or EM leakages. The described technique in this paper specially bolsters the typical dual-rail approach - *Dual-rail Precharge Logic* (DPL) [2], for gaining highly secure dual-rail manners. Generally, DPL employs a *'dual-rail'* and *'dual-phase'* protocol, in which, each logic value a is replaced by a pair of complementary values a_t and b_t respectively in two rails (*True* (T) and *False* (F) rails). The T/F rails work in two alternative phases. In the *'evaluation'* phase, all the effective values are propagated through the combinatorial logic chain, and in the *'precharge'* phase, all the non-register values are reset to a fixed state (normally '0'). A proper realization of this structure ensures only one switch in each clock cycle in view of each compound gate, so as to attain a constant logic behaviour in view of two rails, as seen in Fig. 1.

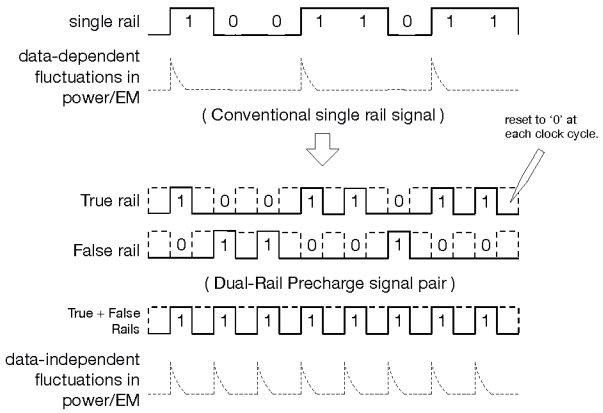


Fig. 1: Dual-rail compensation logic.

To achieve highly complementary dual rail manner, symmetric networks must be achieved to have identical signal propagation and parasitic capacitance. The most widely used technique for this goal is the copy&past by cloning the netlist of the original single rail at the back-end stage to avoid the uncontrollable and unpredictable optimizations during the synthesis and implementation stages. A LUT based compact DPL style, Precharge-Abosrbed DPL (PA-DPL), was described in [3], which aims at achieving very identical dual-rail routings in both separate and interleaved placements. PA-DPL was later found to be vulnerable in Early-Propagation-Effect (EPE) [4] from its second LUT stage. However, a short logic chain by

pipe-lined structure is able to alleviate this defect.

Since the continuously shrinking size of CMOS in deep-submicron technology, *Process Variation* (PV) is posing significant impacts to circuit performances as well as security assurances, which renders PV inevitably become a critical metric to make some design essentials probabilistic and unpredictable [5] [6] [7]. Some researches have certified that silicon process variation brings electrical influences to power consumption from both routings and gates (output capacitive load [8]). This observation also applies to the DPL style, wherein a phenomena arose is that the mismatch between the logic behavior over complementary rails is likely to emit revealable side-channel leakage both on power and EM behaviors. Researches presented in [9] [10] [11] certify that temperature lineally and unidirectionally changes the frequency of RO pairs. This conclusion delivers a possibility to use a thermal generation system to alternatively alter the environmental temperature on-the-fly, for compensating the signal propagation skew in dual-rail format. Benefitted from its highly symmetric rails, the influence from silicon bias can be safely evaluated. In this paper, a system relying on a self-tuned thermal compensation system is described, with the purpose of enhancing symmetric *Precharge Absorbed-DPL* logic by achieving dynamic dual-rail compensation.

The rest paper is organized as follows: Section II gives the prior relevant work; In section III, the proposed thermal system is described over its logic principles; Section IV details the system integration; The security validations from the practical power-based correlation analysis towards a lightweight crypto coprocessor is depicted in Section V; Section VI draws the work conclusions and perspectives.

II. Work Preliminary

A. Technological Process Variation

Process variation is basically an innate phenomenon due to the randomly dispersed articles and etching deviations in the chip fabrication. These unbalanced distributions unfavorably cause tiny spatial differences in electrical characteristics. Since it is deeply rooted at the manufacture stage, to counter its influence in circuit performance become challenging compared to the performance enhancement over logic, behavioral or system level. The influence of process variation to a pair of compensated rails can be observed from two aspects: (a) the symmetric behavior of T/F nets, and (b) the symmetric behavior between each complementary gates as well as their input routings. Many prior work have studied the two aspects within a wide spectrum. In the sequel, an ideally implemented net pair must have the same electrical characteristics, or precisely, their parasitic capacitances must consume the same amount of power in a similar logic behavior. In view of the equation shown in Eq. (1), where P_w denotes the power consumption from a routing when its parasitic capacitance is C_{pc} working under the average flip frequency f and voltage swing V, identical routing length must be maintained for having the same C_{pc}. Moreover, the gate switching actions are affected by the input signals, determined by the arrival time of the effective signals. In this term, the same logic delay in the combinatorial logic chain is necessary.

$$P_w = \frac{1}{2} \cdot C_{pc} \cdot f \cdot V^2 \tag{1}$$

B. Previous Counteracting Approaches

Numerous solutions have been proposed in literatures to alleviate the security defects. A masking solution is stated in [12] to force the interconnect pairs randomly swapped with a bit masking to make the routing bias ignorable. While further investigation reveals that it is not secure enough to overcome the unbalanced routings. A solution described in [13] mitigates the PV by increasing transistor channel length based on *SPICE* simulation. But it requires new fabrication process, and the overall performance in deep-submicron technology inevitably deteriorates. Techniques presented in [14] and [15] both create identical routing networks in interleaved format. This dense placement reduces the relative process variation between each corresponding dual rails, whereas minor technological bias still cripples the security, and the difficulty to make two rail interleaved as well doubts. Technique proposed in [16] purposely serves to minimize the skews between T/F routing delays in a swapped dual-rail format. This technique utilizes an F routing selection algorithm with some unique references. By restricting the candidate F routings with a threshold, a counterpart routing with similar delay (*w.r.t* T routing) can be found. However, the T and F routing pair still poses vulnerabilities over process variations owing to their non parallel paths since the routing pairs are estimated by length, instead of the precise shape. To remove the implementation barriers from mainstream FPGA devices, some expensive efforts must be devoted. Specially devised route techniques presented in [16] [14] [15] partially alleviated this problem in FPGA scenarios, whereas, uncontrollable process variation still causes uncertainties in view of both net and gate pairs.

III. Thermal Compensation System

A. Ring-Oscillator Use in PUF

RO is a low-cost logic that has been widely used in phase-locked loop and Random Number Generator (RNG). The basic structure of RO is a cascaded delay stages in a closed chain, for outputting oscillation with regulated and stable frequency. One of the most important RO usages in security domain is the RO based *Physical Unclonable Function* (PUF) [9], in which, identically laid-out ROs are implemented. Due to the manufacture variations across the chip, each RO delivers slightly different frequencies that cannot be cloned. By deploying a number of identical ROs and comparing the frequencies obtained from the same sequence of RO pairs, a particular output bit combination can be achieved. Importantly, due to the PV from chip to chip, these bit orders are not able to be duplicated even if the internal structure and sequence are disclosed. Hence this logic can be safely used as a unique signature for security authentication.

B. Thermal Influence

Temperature is a major metric that potentially alters the circuit electrical performance owning to its influence to the silicon characteristics. A number of publications have certified that fluctuating local temperature causes frequency deviations for an implemented Ring-Oscillator (RO) [9] [10] [17]. This result implies tiny and unpredictable electrical parameters to a circuit when the environment temperature is swinging or the circuit is heated up by the implemented logic itself [11] [18].

Fig. 2 plots the *Frequency-to-Temperature* relationship between a pair of ROs [9]. The solid and doted lines represent each of the RO pair respectively. Due to the diverse impact from PV over different chip locations, the frequency of the RO represented in dotted line descending faster than that of the RO represented in solid line when the temperature is rising. To a concrete term, temperature alters the circuit frequency inconsistently, *i.e.,* the transistor or routing delays are inversely related to the silicon temperature, because a higher temperature intrinsically reduces the mobility of electrons and holes, which in turn slows down the current speed [19].

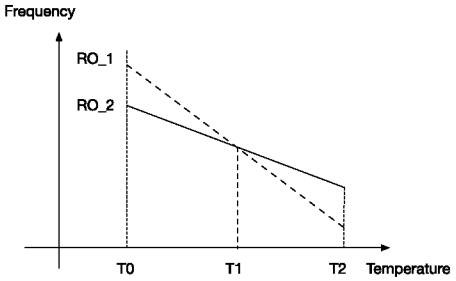


Fig. 2: Frequency-Temperature relationship from a pair of ring-oscillators with different base frequencies.

C. Architecture of Self-Tuned Thermal System

Two groups of heaters are integrated to selectively and dynamically increase the temperature of the location where the corresponding RO has higher frequency in real time. The proposed thermal system inherits the general architecture of RO PUF, but uniquely contains a frequency detection module together with a thermal generation module. A series of inverter based high-frequency ROs play the role as the heater. The prominent heating effect of the 1-invertor RO was described in [20], where authors announced an increased temperature of 135^{o}C using 1,000 LUT based one-stage ROs. Based on our requirements, we used the same 1-inverter RO and utilized a smaller number of ROs to ensure that the local silicon not to be overheated, as described in Fig. 3.

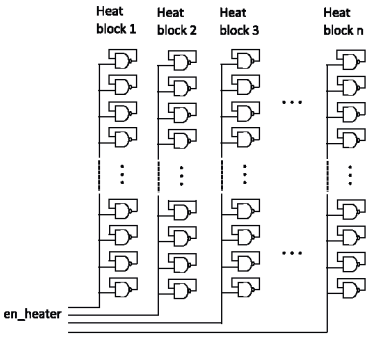


Fig. 3: Network of high-efficient 1-inverter RO-based heater.

The basic structure of the system is illustrated in Fig. 4, which employs two ROs embedded inside each crypto core part as the frequency sensors. Each sensor consists of 15 inverters that yields moderate frequency range to be captured by the counter. The A/B heat generators are deployed closely around

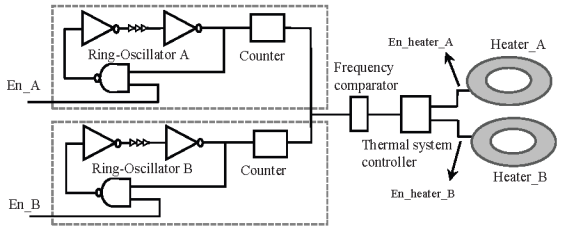


Fig. 4: Architecture of thermal compensation system.

each crypto core for spatially heating up the silicon. The built-in delay cells of the ROs are LUT based inverters in odd number as described in Fig. 4. The number of inverters, logic interconnects and routing delays jointly determine the toggling frequency between '0' and '1'. For exclusively sensing the PV difference, we implemented the A and B ROs by instantiating the same RO hard-macro with fixed routing network. This ensures that the only factor influencing the output frequency is the silicon variation.

D. Thermal Compensation Protocol

The output cycles from each RO is counted inside a fixed time window, and they are compared to determine which RO operates in higher frequency (larger recorded cycle number in this time period). Note that frequency is unidirectionally changed by local thermal environment, *i.e.,* ascending temperature leads to descending frequency, and vice versa. Accordingly, the frequency comparison gets a thermal gradient between the T/F areas. Since a running RO heats the local silicon, a protocol is employed in order to tune the relative frequency difference between the two locations in real-time. The complete functional loop consists of different states. In the initiate state, both A/B ring oscillators begin to work. The frequency of each A/B ROs are computed and measured after a specific time in measure and count states. The surrounding heater in which the RO has higher frequency is switched on for creating higher thermal, which consequently increases the frequency of the influenced RO. Fig. 5 depicts the state transition of the system and its time line for the main states.

E. Gradient Heating

The controller module, working in a state-machine as shown in Fig. 6, does not solely determines which RO runs in higher frequency, but also produces a gradient of the difference. All the 1-inv RO heaters are grouped into numerous heat blocks. Each heat block contains a number of 1-inv RO heaters, and they are controlled by a buffer-driven *enable* signal, which activates different amount of heater blocks according to the counted cycle difference from A and B ROs. The bigger the frequency difference is, the more the heat blocks are activated for generating more heat to the rail with higher frequency. In the sequel, the thermal adjustment in real-time can be smooth without drastic temperature change.

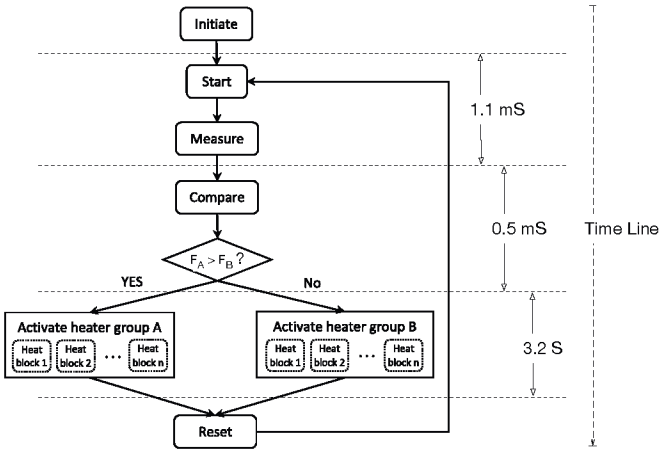


Fig. 5: State flow of thermal compensation.

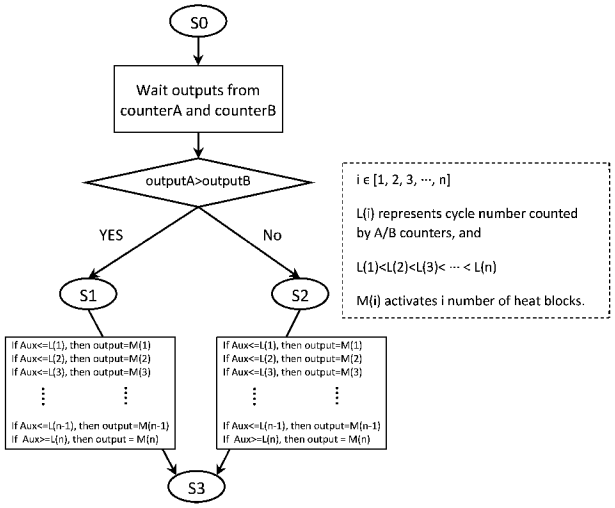


Fig. 6: Regulation of heat blocks for gradient heating.

IV. System Implementation

A. A Case Study

To validate the security level of the devised system, an ultra lightweight secret key cryptography - PRESENT [21] is selected as the case study. *Present* cryptography employs a 64 bit block cipher using either 80 bit or 128 bit key. The basic architecture of *Present* obeys the principle of *Substitution-Permutation Network* [22], which has 32 encryption rounds, consisting of bixor operations between plaintexts and round-keys, a series of nibble-width (4-bit) *S-box* for block substitutions and *pLayer* block for bit permutation. Fig. 7 briefs a parallel architecture of the selected core.

B. Back-End Implementation Flow

The complete design flow of the system consists of reentrant back-end combinations. First, the original PA-DPL dual-rail logic is created obeying the design flow elaborated in [3], and the thermal system is synthesized, placed and routed using commercial tools. Note that the uses of *slice* are free of logic

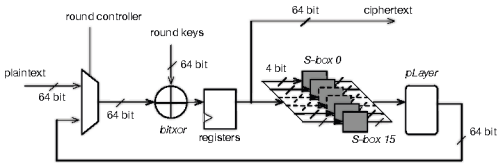


Fig. 7: Parallel Structure of *Present* LightWeight Crypto Core.

overlapping between the two circuits, which is guaranteed by creating placement "*prohibit*" constraint in the user constraint file. Finally, the dual-rail crypto design and thermal system design are merged at XDL stage. More exactly the crypto cores are boxed with the heater belt, as illustrated in Fig. 8. Moreover, 100, 140, 180, 220 1-inv heaters are activated respectively for 4 gradient heating grades.

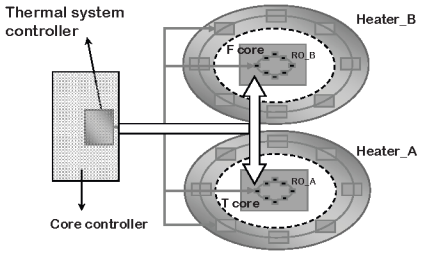


Fig. 8: Implementation of thermal system embedded dual-rail crypto core.

For safely implementing the ROs with smaller base frequency difference, a surface scanning is launched before determining the locations for deploying the T/F crypto cores. Table. I shows the surface scanning result by identically situating a RO hardmacro primitive (inverter loop with 31 inverters) into different clock regions, to have *identical* logic configuration and routing paths. The counted cycles at a fixed time window differ clearly because of the process variation. More precisely, the RO frequencies are generally descending from clock region 1 and 2 (south side in *FPGA editor* view) to clock region 11 and 12 (north side) in this tested FPGA. It is emphasized that the PV distribution from chip to chip might vary significantly, so it is safer to have this process done before the implementation, so as to ensure that the thermal introduced is sufficient to alternatively switch the frequencies of the two ROs. Fig. 9a and Fig. 9b shows the dual core structure of *Present* in separate placement and in row-crossed interleaved format. SASEBO-GII [23] is chosen as the implementation platform, which contains a Spartan-3 FPGA performing as the controller to feed plaintext and receive ciphertext from the crypto core implemented on the main Virtex-5 FPGA.

V. Security Evaluation

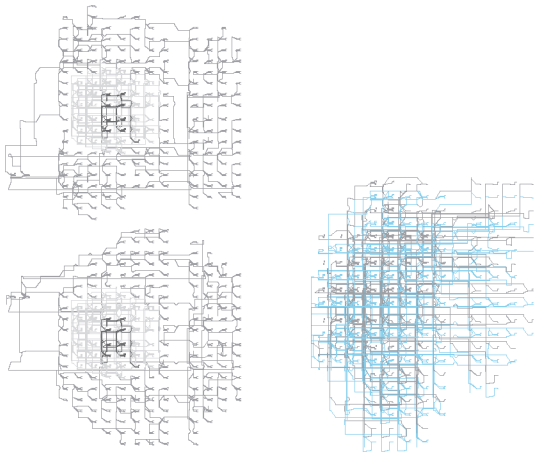
A. Comparison Implementation and Analysis Model

To evaluate the effectiveness of the devised compensation systems, we have conducted comparison attacks using the generic correlation model. Three main counterparts are prepared: (a) Circuit a is a single-rail *Present* crypto core directly implemented over Virtex-5 FPGA without special placement

TABLE I: PVs Influenced Ring Oscillator Output

Region	1	3	5	7	9	11
Cycle No ($\times 10^3$)	*10.89*	*10.78*	*10.68*	*10.58*	*10.53*	*10.47*
Region	2	4	6	8	10	12
Cycle No ($\times 10^3$)	*10.86*	*10.76*	*10.70*	*10.64*	*10.53*	*10.52*

A 31-stage RO primitive is tested on Virtex-5 FPGA (XC5VLX50), consisting of 12 clock regions. Time window for cycle counting is $200\mu S$.



(a) Circuit A (b) Circuit B

Fig. 9: Circuit A: Separate dual-rail *Present* cryptography with thermal system; Circuit B: Interleaved dual-rail *Present* with identical routing networks.

constraints; (b) Circuit b is dual-rail circuit transformed from Circuit a, combined with the proposed thermal system; and (c) Circuit c is the dual-rail one manipulated with interleaved placement, as seen in Fig. 9b. To validate the impact of heating to SCA-hardened DPL, circuit b is attacked twice, with thermal system on and off respectively. The target logic point here is the last round 64-bit registers, for disclosing the 64-bit last round key. Since *Present* cryptography employs 16 4-bit data/key blocks so each attack is actually retrieving a 4-bit (2^4=16 candidate) subkey. The targeted logic point is the register cluster at the last computation round using *Hamming Distance* model, as seen in Fig. 7. The inverse *pLayer* and *S-box* operations are executed in the prediction model. Because the permutation operation is after the *S-box*, the selected 4 key bits should be mapped to the same *S-box* in the last round. Eq. (2) gives the power hypothesis function for the 4-bit registers of each *S-box*, which corresponds to different key bits and ciphertext bits mapped by $pLayer$ operation $P_{(i,i+1,i+2,i+3)}$ [21].

$$Power_{(i,i+1,i+2,i+3)} = \{S^{-1}(C_{P_{(i,i+1,i+2,i+3)}} \oplus K_{P_{(i,i+1,i+2,i+3)}}) \oplus C_{(i,i+1,i+2,i+3)}\} \quad (2)$$

B. Result Analyses

The *Present* coprocessor is controlled by a terminal in PC, which is synchronized with the *OS* in Lecroy Oscilloscope (*Waverunner 610Zi*) for sending plaintexts from a *PRNG* and storing the measured power traces. To facilitate the data transmission to the workspace, segmented memory is activated in Oscilloscope with 5,000 pieces. Every aligned trace for each encryption is stored in one piece of memory, and totally 10 groups of traces with different plaintext are running repetitively, yielding a ciphertext quantity of 50,000. To minimize the environmental noise, we have repeated the same trace measurement for 100 times, and the traces for the same plaintext encryptions have been averaged.

Tab.II shows all the restored subkey nibbles from the comparison attacks. For the unprotected single-rail one (circuit_*i*), all subkeys are successfully cracked. For the dual-rail one with thermal system switched off (circuit_*ii*), 6 out of 16 subkeys are recovered. For the dual-rail one with thermal system on (circuit_*iii*), none is succeed. In contrary, one subkey is recovered for the interleaved dual-rail circuit (circuit_*iv*). Many factors potentially affect the results, hence the revealed key nibbles for each time might differ in each test. To sketch a general trend, we plotted the *right key rank position* to check whether the correlation peak for each key nibble is closer to be revealed among the 16 candidate nibbles, as plotted in Fig. 10. It can be seen that except the single-rail one, the one with heater system off is closer to be fully discovered. The one with heater on and the one with interleaved placement have lower rank positions, *i.e.,* right subkey nibbles are more difficult to be differentiated out. Moreover, the circuit_*iv* has similar result due to its symmetric and interleaved networks that make its accumulated PVs for the paired rails very small. Since this circuit can only be achieved by manipulation of XDL that is not supported by new device (Xilinx 7 series FPGAs). Technique in this paper provides an alternative with competitive security grade for the latest devices.

TABLE II: Comparison Attack Results

Last round key after inverse *pLayer* permutation: *'5C3C8FEF91F30FF2'*

subkey (i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
revealed	***5***	***C***	***3***	***C***	***8***	***F***	***E***	***F***	***9***	***1***	***F***	***3***	***0***	***F***	***F***	***2***
subkey (ii)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
revealed	*1*	***C***	*B*	***C***	***8***	*7*	*B*	***F***	***9***	*4*	*5*	*B*	*E*	*C*	***F***	*4*
subkey (iii)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
revealed	*C*	*1*	*6*	*5*	*E*	*4*	*B*	*B*	*6*	*3*	*A*	*8*	*6*	*7*	*7*	*D*
subkey (iv)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
revealed	*9*	*8*	*4*	*B*	*7*	*7*	*E*	*9*	*B*	*A*	*A*	*7*	*5*	*6*	*A*	*B*

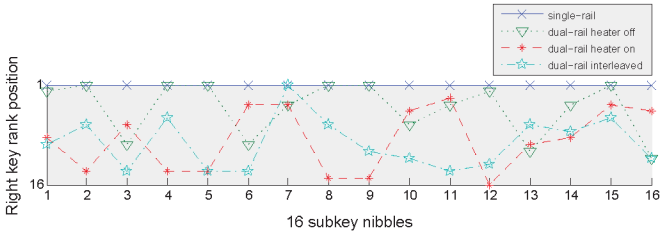


Fig. 10: Rank position of the right subkey correlation peaks.

C. Further Discussion

The proposed thermal system might not be suitable to be directly applied to a large crypto algorithm. This is because

the implementation area is likely to have more significant dispersion of process variation, and the heater hence cannot sufficiently alter the thermal environment to all the security-sensitive logic elements. However, a careful logic partition can spatially concentrate to a smaller security-critical area. It is emphasized that the changing heating pattern leaks some internal information, which might be exploited by adversaries to make pattern profiling. However, this pattern only reflects the averaged effect of the temperature over the entire chip and dynamic switches make the temperature change in a non-linear process, which does not directly or intermediately response to the calculated single bits. In addition, the heating behaviour is unpredictable for every chip due to the unique technological bias. In this term, the template attack cannot be applied, which increases the complexity of the potential attacks.

VI. Conclusions

In this paper, a self-tuned thermal compensation system for alleviating the impacts of silicon process variation in SCA-resistant dual-rail logic is described. This system relies on a solid thermal effect over trivial silicon technological bias to selectively affect the parasitic parameters between the complementary T and F rails. Identical ROs are embedded inside each dual-rail parts, playing the role as the thermal sensor, and outputs the counted cycles from the two ROs. Frequency comparison after a fixed time window determines the faster RO to be heated up for slowing down its frequency in automatic manners. The main benefit of this system lies within its automatic adaptation to the silicon thermal environment on-the-fly, to be amended automatically by alternatively heating the crypto core where the faster RO resides. Experimental attacks demonstrated that by tuning on the thermal compensation system the security level of the dual-rail circuit is increased significantly and competitive to the interleaved dual-rail circuit. Since the implementation process is free of the bottom-layer XDL manipulation, the proposed system is fully adaptive to the newest FPGA devices.

In the subsequent work, we will interleave the heat system with the crypt cores in order to attain better thermal effect using smaller heat expenditure.